基于元搜索的网页消重方法研究

Study on the Duplicated Web Pages Detection Algorithm with Meta Search Engine

谢　蕙　秦　杰　（河南工业大学 信息科学与工程学院 河南郑州 450001）

摘　要：本文在对现有主流网页消重技术进行分析基础上，针对元搜索引擎技术，提出一种基于元搜索的网页消重算法。介绍了算法的具体实现步骤，并且通过实验验证了算法的有效性。

关键词：消重 特征码 元搜索引擎 网页元数据 摘要相似度

1 引言

随着网络技术的飞速发展，网络提供给人们的信息资源越来越多，要想在浩瀚的网络资源中查找到有用的信息，搜索引擎起到了重要作用。但是现在的搜索引擎技术并不完善，存在一些亟待解决的问题，最主要的问题之一是存在大量的重复网页[1]。

对用户而言，如果查询到的是重复信息，严重影响查询效率。对互联网系统而言，如果采集到大量重复网页，既浪费信息检索时间又浪费存储空间。网络机器人（spider）采集互联网中的相关信息，采集信息的数量非常巨大，采集结果的处理，需要进行消重处理后，才能作为有效的信息。若单单依靠人工进行去重处理，不仅耗费宝贵的人力资源，而且时效性也不能满足实际需要。

为了解决这些问题，网页消重成为搜索引擎所研究的一项重要技术。

2 网页消重技术

网页消重技术是指对内容重复的网页进行识别，处理和合并，以节省网页数据库的存储空间和在网页数据库上进行操作的时间的过程[2]。

2.1 网页消重技术主要思想

根据网页重复原因不同对应的判断网页是否重复的方法也有所不同，但是几乎所有的消重技术都基于这样一个基本思想[3]：根据一定的算法为每个文档计算出一组指纹（fingerprint），若两个文档拥有一定数量的相同指纹，则认为这两个文档的内容重叠性较高，也即二者是重复网页。

对于 URL 相同的网页，为了避免重复搜集同一 URL 网页，主要使用排除相同 URL 的方法：网络蜘蛛把访问过的网页地址变成信息指纹存放到哈希表中，在抓取新的网页时，把网页的地址解析成指纹，判断比较该指纹是否存在于哈希表中，若存在则表示已下载过，若不存在则下载且把这个指纹存放到该哈希表中。当然这个算法要保证几乎不能产生相同的网址指纹。

针对内容相同的网页，当前比较成功的搜索引擎系统大多是基于关键词匹配和结合向量空间向量模型来完成消重任务的。典型的系统包括 Google 和天网系统[3]。通常这类系统的消重工作过程大致是：在对网络蜘蛛已抓取回来的网页进行分析时，首先对网页进行净化，提取出网页的主题以及与主题相关的内容，这些内容包括网页标识、网页类型、内容类别、标题、关键词、摘要、正文、相关链接等信息，根据网页的关键词、摘要、正文等信息提取网页的特征项，构造评价函数，根据两个网页的特征项的相似度判断网页是否重复。

2.2 现有主流网页消重技术[1]

SCAM 算法计算出每篇文档中各个单词的词频，将文档用词频向量的方法表示出来，计算 2 个词频向量之间的距离，在一定的范围之内就判断为相似的文档。DSC（digital syntactic clustering）算法，首先将一篇文档分成由 n 个字组成的 shingles，一篇文章就可以由 n 个 shingles 来表示，再按照一定的过滤规则将过滤出

的 shingles 作为该篇文档的代表,参加比较的就是这些被选出来的 shingles。I – Match 算法是对 DSC 算法的一种改进,它从过滤 shingles 这方面着手,尽量过滤掉尽可能多的重复次数较多的 shingles。

北大的天网系统在搜集并分析一篇网页时,提取并记录了网页中出现的关键词,同时根据公式赋予每个关键词一个权值,这些关键词的权值构成一个向量空间,可以用来表示该网页。并以关键词作为网页的特征项。清华大学使用的提取方法是在文章中逗号,句号的前后各取 2 个汉字,作为字符串。哈工大使用的方法是在文章中各个句号的前后各取 5 个汉字。虽然提取汉字的方法不同,但是都是以标点作为文中的提取标记,这种方法效率较高,因为提取字符串是线性时间的,就是把一个 $O(n^2)$ 时间复杂度的问题转变成了 $O(n)$ 时间复杂度的问题,不失为一种好方法。

2.3 现有网页消重技术的局限

我们为网页消重算法设计的评价指标包括算法复杂度、查全率和准确率三个方面,其中查全率是指去重算法所发现的转载网页占总网页的百分比,而准确率反映了算法所发现的转载网页中有多少是真正的转载网页。

当前提出的网页消重算法还比较少,在这些算法中还存在着很多的局限。

SCAM 算法去重效率不高,要求存储空间较大;DSC 算法准确率不高,比较次数过多,效率下降;I – Match 算法效率和准确率比较平均,占用硬盘空间大。国内几种算法大都是对国外算法的沿用,在算法效率,准确率,时效性等方面都存在各种问题。

3 一种基于元搜索引擎的网页消重方法

该元搜索引擎模型,选择若干成员搜索引擎,针对用户的单个查询请求,调用成员搜索引擎的搜索结果,存储到数据库,经过相应的结果融合,再返回给用户。

不同搜索引擎的检索结果中会有一定程度的重复,为了使元搜索引擎获得用户满意的高质量的检索结果,必须尽量消除重复。一般来说,会有以下几种情况:①最简单的重复情况是结果具有相同的 URL,可以很容易的排除;②同一文档存在常见的别名,或者是同一文档被做了链接因而具有差异较大的别名;③同一文档具有不同的版本或拷贝,放在不同的位置,此时存放的主机也可能不相同,此种情况最难识别。

考虑到元搜索引擎的特殊性,可以充分利用成员搜索引擎提供的结果基本信息,如网页地址,网页标题,网页摘要等。因此,该网页消重算法选择结果网页集合中每条记录的网页地址,网页标题,网页摘要作为判断网页重复性分析的依据。算法具体设计方案如下:

(1) 网页元数据提取

元搜索引擎模型调用源搜索引擎,得到搜索结果——结果网页记录集,直接提取网页集合中每条记录的标题,地址和摘要作为网页元数据。

(2) 提取网页摘要特征串

针对网页的摘要,用文本中分隔标记把文本分成若干句子,从句子中提取特征码,把特征码按顺序连接起来构成该文本的特征串。

(3) 摘要相似度评价函数

为了实现去重模块中判断摘要相似度功能,摘要相似因子设计如下:

FazzyFactor = pn/wn。其中 FazzyFactor 是相似因子,pn 是两个特征串中特征码相同的个数,wn 是两个特征串的特征码的个数,相似因子的值即为相似度。

判断 wn 的值:如果两特征串含特征码的个数相等,则 wn 的值即特征串的特征码个数值;否则是两个特征串的特征码个数的较小值。

判断 pn 的值:比较特征码是否相同。每有一组特征码相同,pn 的值就加 1。

设计系统阈值 OF,若两个摘要的相似因子小于该阈值则该两个摘要重复,否则不重复。

(4) 消重算法描述

① 提取记录的网页元数据,判断网页地址是否重复。如果地址相同,则重复,转⑥;否则,转②;

② 判断网页标题是否重复。如果标题相同,则转③;否则,转④;

③ 依次提取每个网页摘要的特征码信息,将提取出的特征码与平衡搜索树中的特征码相比较,判断相似度,若相似度大于系统阈值 OF1,则两条记录重复,否则,转⑤;

④ 依次提取每个网页摘要的特征码信息,将提取出的特征码与平衡搜索树中的特征码相比较,若相似

度大于系统阈值 OF2，则两条记录重复，否则，转⑤；

⑤ 将提取出的特征码插入平衡搜索树中，并转①；

⑥ 结束。

4 实验验证

采用以上介绍的算法，我们在元搜索引擎系统中对一批数量在 100—200 的网页集合进行处理，将实验结果与人工判别的结果进行比较，发现重复网页的准确率达到 96 %以上。

在成员搜索引擎个数固定的情况下，我们对算法的响应时间做测试，测试结果如表 1 所示。从实验结果可以看出，去重处理过程中的主要时间用于特征码的提取。

表 1　算法去重处理时间

网页数目(个)	特征码提取时间(s)	去重处理时间(s)
100	<1	1.2
500	1	1.5
1000	1.2	1.8
1500	1.6	2.2
2000	2.1	3

当结果集合网页数目固定时，我们对算法执行时间与成员搜索引擎数目的关系做了测试。测试结果如图 1 表明，选择适当的成员搜索引擎，权衡它们的数量和性能，才能充分发挥该算法的性能。

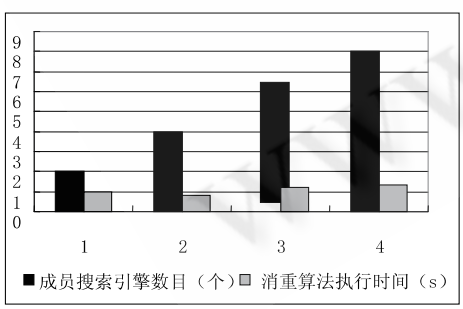

图 1　算法性能与成员搜索引擎个数的关系

5 结束语

将该方法用于元搜索引擎系统中，能有效提高检索质量，返回给用户更为准确的结果。

经分析发现，这种算法的主要缺陷在于所提取的特征码信息只代表了这些标点周围的信息，未提供网页摘要的内容信息。并且，算法的去重过程中主要时间用于特征码的提取。所以下一步工作是重点改进特征码提取方法，提高特征码提取效率并且使特征码更能表现网页摘要的内容。

参考文献

1 白广慧. 网页排重技术研究与应用. 中国科学院，2006.

2 陈基漓，牛秦洲. 基于特征码的网页去重. 微计算机信息，2006，22(3)：113－115.

3 吴平博，陈群秀，马 亮. 基于特征串的大规模中文网页快速去重算法研究. 中文信息学报，2003，17(2)：28－35.

4 J. Zhou，P. Larson，J. C. Freytag，W. Lehner. Efficient Exploitation of Similar Subexpressions for Query Processing. ACM SIGMOD，2007：533－544.

5 郑德全，胡熠，于浩，赵铁军，王青松. 多载体数据流中的特定信息识别研究. 软件学报，2003，14(9)：1538－1543.

6 Junghoo Cho，N. Shivakumar et al. Finding replicated web collections. In Proceedings of 2000 ACM International Conference on Management of Data (SIGMOD)，May 2000.

7 Shaozhi Ye，Ji－RongWen，Wei－Ying Ma. A systematic study on parameter correlations in large－scale duplicate document detection. Knowledge and Information Systems，2007，14：217－232.